Robust change-point estimator for series of normal observations.

Alvaro Eduardo Cordero-Franco, Jorge Arturo Garza-Venegas
Universidad Autónoma de Nuevo León
San Nicolás de los Garza, Nuevo León, México

Víctor G. Tercero-Gómez
Tecnológico de Monterrey
Monterrey, Nuevo León, México

Abstract

In statistical process control, detection of special causes of variation and the estimation of the time when they occur are two important tasks for process improvement. When dealing with normal independent observations, maximum likelihood estimators for a change-point have been derived, and their structure happens to correspond with a least squared formulation. This formulation facilitates the estimation of a change-point. However, in the presence of outliers, the estimation bias increases. To deal with this problem, regression analysis uses, among other procedures, a technique called least absolute residuals, where the norm $L_2$ is replaced by the norm $L_1$ in the objective function to be minimized. Using this idea, a robust estimator for a change-point in a series of independent normal observations is developed. Preliminary simulation performance results are presented and compared with the corresponding maximum likelihood estimators in the presence of outliers. Additionally, practical implementation is described by using a numerical example. It is expected that this estimator is found useful for practitioners when dealing with suspicious outlying observations.

Keywords
Change-point analysis, MLE, Least Absolute Deviations Estimator, Outliers.

1. Introduction.

In Statistical Process Control, observations from a process are analyzed in order to determine whether or not its behavior (by means of variability) is under statistical control or not; and if it is not in control, corrective actions has to be taken to secure a state of statistical control. This task is done as follows: first, a monitoring stage, which is done mostly with control charts (CC); after the CC signals an out of control process, then a corrective stage takes place, in which some actions need to be taken to get the process back to statistical control; and the cycle begins again. Control charts are statistical tools to monitor processes, but they cannot determine accurately the initial moment when the process get out of control. Knowing this moment has positive consequences as reduction of costs of rework, trash and stopping work lines (time for searching the root cause of the change) and boost of the improving stage, among others.

Change-point problem is known as the problem to detect and estimating the initial moment when a process changes, while the collection of methods address this issue are called change-point analysis (CPA). Besides CC’s, there are many tools that have been developed to have an estimation of the change-point, and they can be used in a complementary way: CC’s to detect out of control processes, and CPA to make the estimations. Amiri and Allahyari [1] summarized works about CPA categorizing them in: approach of solution, type of change considered, control chart used (if it is the case), etc.

An outlier is an observation which appears to be far from the other data. In practice, it could happen that a measure of some characteristic is registered wrong, because of tool’s calibration or typographical error. These errors necessarily affect the estimation in some way. Use of least absolute deviation estimators in robust regression is
useful when dealing with atypical observations. Dielman [2] showed that these estimators present better results in terms of bias and standard error compared against least squares estimators.

This work presents a change-point estimator based on the use of norm L1 in a normal observation series with a shift in the mean. This estimator is developed in order to analyze its behavior when presence of outliers is suspected. This work shows which estimator (the one based in the norm L2 against the one based in the L1 norm) is recommended to use when presence of outliers is suspected, and therefore the robustness to outliers of a change-point estimator. In order to do that, normally independent distributed observations with a shift in the mean are considered and outliers of different magnitude will be added to the original series. MLE (norm L2) and LAD (norm L1) estimators are presented as well as a numerical example.

2. Previous work.

Change-point analysis started from a Bayesian point of view with Girshick and Rubin [3]. They defined a quality control rule to detect a change in a random process. Page [4-6] from a frequentist point of view, developed a control chart based on cumulative sums: the cumulative sums (CUSUM) chart, which in presence of sustained changes are faster to detect than traditional Shewhart’s control charts [7].

Hinkley [8] set the basis to construct maximum likelihood estimators (MLEs) and likelihood ratios tests (LRTs) from a parametric approach. These techniques were applied to independent and normally distributed observations. Following this guideline, MLEs for others distributions were developed by some authors (Samuel, Pignatiello, and Calvin [9,10], Samuel and Pignatiello [11], Dabye and Kutoyants [12]). Not only MLEs were developed in order to estimate the change-point in a time series, for instance, the technique proposed by Page [4] for his CUSUM chart could also provide an estimation of the change-point which was later compared with the MLE of Hinkley [13] in 1971, who conclude that CUSUM is asymptotically biased, but easier to use. This MLE was developed for normally distributed series assuming prior knowledge of initial parameters. Using EWMA charts and Moving Average charts change-point estimators could also be obtained, and Nishina [14] compared its performance against CUSUM estimators, determining the superiority of the CUSUM and showing that CUSUM estimation is an MLE when there is prior knowledge about the shift in the mean. Recently, Tercero et al. [15] proposed a likelihood ratio test for shifts in the mean of normal independent time series, providing the numerical quantiles of the test statistic for different significance levels.

On the other hand, from a non-parametric approach, Page [5] proposed a method using the sign function to determine changes in the mean of symmetrical distributions. Later, Bhattacharyya and Johnson [16] developed a test assuming symmetry of the cumulative distribution of the observations. Pettitt [17] worked with Bernoulli, binomial and continuous observations proposing a nonparametric test for the presence of a change-point. Bootstrap technique was used by Hinkley and Schechtman [18] to compare both parametric and non-parametric methods for shifts in the mean. In 2007, Zou et al. [19] solve the issue of the estimation of the change-point and the estimation by proposing the use of the empirical likelihood ratio. Recently, Tercero et al. [20] assessed the problem of estimating changes in time series using p-values of a non-parametric test: Mood’s median test.

In a parametric approach, the change-point estimation may present a higher bias in presence of outliers and in normal processes change-point MLE has the form of least squares estimator (LSE). For parameter’s estimation in regression models besides of LSE, least absolute deviations estimators (LAD) are used to provide better estimations in terms of bias and standard deviations in presence of outliers. This method estimates regression parameters by minimizing the sum of the absolute deviations from the mean; that is, using minimizing L1 norm. Boscovich [21] first discussed the minimization of the sum of the absolute errors to fit a line to observed data. After this work, many authors have analyzed the use of LAD for different regression models and different approaches of solution of the minimization problem implied in least squares estimation. Dielman [22] listed articles of LAD estimation applied to both linear and non-linear statistical regression models. Powell [23] proposed a generalization of LAD estimator for linear models, which is consistent and asymptotically normal distributed for a wide class of errors’ distribution and robust to heterocedasticity. Dielman [2] showed that LAD estimators are superior in their performance to LSE estimators in presence of outliers in linear regression models. Pollard [24] proved the asymptotic normality of LAD regression estimators. Morgenthaler [25] analyzed the consequences of using LAD estimators instead of LSE ones in linear regression models. Bai [26] obtained rates of convergence and asymptotic distribution for LAD parameters in linear regression. Dielman [27] presented a literature review in LAD regression. No LAD change-point estimators were found in literature.
This paper considers the estimation of a change-point in a series of independent and normal observations in the presence of outliers. Change-point MLE and LAD estimators are presented and their performance is compared by using simulation and considering different scenarios.

3. Problem statement.
Suppose a series of $T$ normal and independent observations at which a shift in the mean have occurred at moment $\tau$ and with parameters before and after that are unknown. This can be expressed mathematically as follows:

$$X_i \sim \begin{cases} N(\mu_0, \sigma), & 1 \leq i \leq \tau \\ N(\mu_1, \sigma), & \tau < i \leq T \end{cases}$$  \hspace{1cm} (1)

Now, suppose that observation $k$ (where $1 < k < \tau$) is an outlier. This research aims at comparing the performance of the change-point MLE against the LAD estimator in presence of outliers by considering different scenarios. MLE and LAD change-point estimators are described in section 4.1 and 4.2, respectively. Section 5 shows a numerical example with the design of the simulation in detail. Conclusions and future work are left in Section 6.


4.1. Change-point maximum likelihood estimator.
Retrospective analysis makes use of all available information as MLE’s. That is why maximum likelihood approach was selected in order to estimate the change-point. For the problem stated, likelihood function is presented in equation (2), while MLE’s for the unknown parameters are presented in equations (3)-(5). It can be seen that the problem grows in complexity because all MLE’s depend of the change-point location. Tercero et al. [28] show a more detailed derivation of these estimators.

$$L(\mu_0, \mu_1, \sigma \mid \bar{x}) = \prod_{i=1}^{\tau} f(\mu_0, \sigma \mid x_i) \cdot \prod_{i=\tau+1}^{T} f(\mu_1, \sigma \mid x_i)$$  \hspace{1cm} (2)

$$\hat{\sigma}^2(\hat{\mu}_0, \hat{\mu}_1) = \frac{\sum_{i=1}^{\tau} (x_i - \hat{\mu}_0)^2 + \sum_{i=\tau+1}^{T} (x_i - \hat{\mu}_1)^2}{T}$$  \hspace{1cm} (3)

$$\hat{\mu}_0 = \frac{\sum_{i=1}^{t} x_i}{t}, \quad \hat{\mu}_1 = \frac{\sum_{i=\tau+1}^{T} x_i}{T-t}$$  \hspace{1cm} (4)

$$\hat{\tau}_{MLE} = \operatorname{arg} \min_{2\tau \leq t < T-2} \{\hat{\sigma}(\hat{\mu}_0, \hat{\mu}_1)\}$$  \hspace{1cm} (5)

Since this estimator considers squared deviations from the mean, it gives more weight to outliers which might create a bias in the change-point estimation. In next subsection an estimator that uses norm L1 is presented providing less weight to outliers. It is based on sums of absolute deviations

4.2. Change-point least absolute deviations estimator.
Instead of minimizing the sums of the squares of the deviations around the mean, the least absolute deviation method considers as estimator the point which minimizes the sums of the absolute deviations from the mean. This method is used in regression, and it is also called the minimum L1-norm regression. Due to this method considers absolute deviations instead of squares; it gives less importance to outliers than least squares method. LAD estimator is presented in equation (6), where the estimators of the mean are the same as in (4).

$$\hat{\tau}_{LAD} = \operatorname{arg} \min_{2\tau \leq t < T-2} \left\{ \sum_{i=1}^{t} |x_i - \hat{\mu}_0| + \sum_{i=\tau+1}^{T} |x_i - \hat{\mu}_1| \right\}$$  \hspace{1cm} (6)
To compare these estimators, a numerical example is presented in next section followed by the analysis of several scenarios considering different factors, such as, size of series, change-point location, shifts in the mean measured in standard deviations, etc.

5. Simulation

5.1 Simulation design

To compare performance of \( \hat{\tau}_{\text{MLE}} \) and \( \hat{\tau}_{\text{LAD}} \), several scenarios were simulated by considering the following factors and levels:

1. Shift in the mean (\( \delta \)) was measured in standard deviations from an initial mean. Shifts considered were 0.5\( \sigma \), \( \sigma \), 1.5\( \sigma \), 2\( \sigma \) and 3\( \sigma \).
2. Series length (\( T \)) of 20 and 50 observations were simulated.
3. Change-point position (\( \tau/T \)) of 0.3 and 0.5.
4. Units (\( r \)) added to an observation to generate an outlier of 0, 4, 6 and 9.

Monte Carlo experimentation was used to evaluate the performance of the change-point estimators. The following general procedure was used for each scenario evaluated:

1. Select the scenario and generate \( T \) random variables by combining factors and levels shown above.
2. Replace \( x_j \) where \( j = \lceil \tau/2T \rceil \) (ceil of \( \tau/2T \)) for \( x_j = x_j + r \)
3. Calculate \( \hat{\tau} - \tau \) (using estimators under analysis).
4. Repeat step 2 and 3 for all values of \( r \) 1,000 times.
5. Calculate the mean and standard deviation of the error.
6. Return to step 1 and select another scenario.

For example, consider a series of size \( T = 100 \) following (1), see Table 1, in which a shift in the mean \( \delta \) of 1 standard deviation has occurred at observation 50. The original in control observation 25 was replaced by making \( x_{25} = x_{25} + r \) over different values of \( r \); these observations are outliers. Table 2 shows the bias of the estimations.

<table>
<thead>
<tr>
<th>Time</th>
<th>Observation</th>
<th>Time</th>
<th>Observation</th>
<th>Time</th>
<th>Observation</th>
<th>Time</th>
<th>Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.64002</td>
<td>21</td>
<td>-0.27712</td>
<td>41</td>
<td>-0.95842</td>
<td>61</td>
<td>-0.75167</td>
</tr>
<tr>
<td>2</td>
<td>0.38336</td>
<td>22</td>
<td>-0.95011</td>
<td>42</td>
<td>1.39348</td>
<td>62</td>
<td>0.82692</td>
</tr>
<tr>
<td>3</td>
<td>1.25172</td>
<td>23</td>
<td>-0.29509</td>
<td>43</td>
<td>0.20526</td>
<td>63</td>
<td>-1.32114</td>
</tr>
<tr>
<td>4</td>
<td>0.31622</td>
<td>24</td>
<td>-0.37171</td>
<td>44</td>
<td>1.94067</td>
<td>64</td>
<td>-1.33496</td>
</tr>
<tr>
<td>5</td>
<td>-0.39891</td>
<td>25</td>
<td>-0.74274</td>
<td>45</td>
<td>1.63880</td>
<td>65</td>
<td>0.75562</td>
</tr>
<tr>
<td>6</td>
<td>-1.05063</td>
<td>26</td>
<td>-0.15499</td>
<td>46</td>
<td>-0.04173</td>
<td>66</td>
<td>1.53653</td>
</tr>
<tr>
<td>7</td>
<td>-2.00558</td>
<td>27</td>
<td>0.79465</td>
<td>47</td>
<td>0.79859</td>
<td>67</td>
<td>2.49463</td>
</tr>
<tr>
<td>8</td>
<td>-0.63191</td>
<td>28</td>
<td>-0.98611</td>
<td>48</td>
<td>0.03480</td>
<td>68</td>
<td>-0.43289</td>
</tr>
<tr>
<td>9</td>
<td>-0.71418</td>
<td>29</td>
<td>0.43534</td>
<td>49</td>
<td>0.65194</td>
<td>69</td>
<td>1.19627</td>
</tr>
<tr>
<td>10</td>
<td>-0.25317</td>
<td>30</td>
<td>1.30972</td>
<td>50</td>
<td>0.38448</td>
<td>70</td>
<td>1.71563</td>
</tr>
<tr>
<td>11</td>
<td>-0.23487</td>
<td>31</td>
<td>-0.68313</td>
<td>51</td>
<td>0.80680</td>
<td>71</td>
<td>0.90275</td>
</tr>
<tr>
<td>12</td>
<td>0.20799</td>
<td>32</td>
<td>0.70660</td>
<td>52</td>
<td>0.47758</td>
<td>72</td>
<td>-0.07057</td>
</tr>
<tr>
<td>13</td>
<td>-0.19263</td>
<td>33</td>
<td>-0.09300</td>
<td>53</td>
<td>0.88489</td>
<td>73</td>
<td>2.39464</td>
</tr>
<tr>
<td>14</td>
<td>-0.88924</td>
<td>34</td>
<td>-0.61656</td>
<td>54</td>
<td>0.96372</td>
<td>74</td>
<td>1.63754</td>
</tr>
<tr>
<td>15</td>
<td>-0.43754</td>
<td>35</td>
<td>0.59520</td>
<td>55</td>
<td>-1.61580</td>
<td>75</td>
<td>0.50358</td>
</tr>
<tr>
<td>16</td>
<td>-0.12969</td>
<td>36</td>
<td>0.76222</td>
<td>56</td>
<td>-0.27454</td>
<td>76</td>
<td>1.39606</td>
</tr>
<tr>
<td>17</td>
<td>-1.82023</td>
<td>37</td>
<td>-0.70047</td>
<td>57</td>
<td>0.59291</td>
<td>77</td>
<td>2.94384</td>
</tr>
<tr>
<td>18</td>
<td>-0.01720</td>
<td>38</td>
<td>-0.50690</td>
<td>58</td>
<td>1.45865</td>
<td>78</td>
<td>2.45983</td>
</tr>
<tr>
<td>19</td>
<td>0.53204</td>
<td>39</td>
<td>1.88609</td>
<td>59</td>
<td>0.80915</td>
<td>79</td>
<td>0.17293</td>
</tr>
<tr>
<td>20</td>
<td>1.13298</td>
<td>40</td>
<td>-0.50346</td>
<td>60</td>
<td>-0.71543</td>
<td>80</td>
<td>0.77966</td>
</tr>
</tbody>
</table>

Table 1. Normal standard observations following (1) with \( \tau = 25 \), and \( T = 100 \). 25th observation of Table 1 is replaced by adding \( r \) to 25th observation.
In this example there is a difference in bias when the outlier is far from in-control observations by 4 to 6 units, and is the same when outliers are far from 9 to 12 units. Figure 1 shows time series with outlier over different values of $r$.

### 5.2 Results

Bias and standard error of change-point estimators are presented in Tables 3 and 4 over scenarios previously described. 1000 replicates obtained from standard normal series were used to analyze performance of MLE and LAD estimator over different shifts in the mean, change-point location and the magnitude of outliers.
It can be seen in Tables 3 and 4 that despite of the size of the series and the magnitude of outliers under analysis, \( \hat{\tau}_{LAD} \) has smaller bias than \( \hat{\tau}_{MLE} \) when change-point is located at the middle of the series. When change-point is located closer to the beginning of the series, \( \hat{\tau}_{MLE} \) has smaller standard deviation almost in all cases over different shifts in the mean. Nevertheless, performance of both estimators becomes similar as shift in the mean increases.

<table>
<thead>
<tr>
<th>( \delta )</th>
<th>( \tau / T = 0.3 )</th>
<th>( \tau / T = 0.5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tau / T = 0.3 )</td>
<td>| |</td>
<td></td>
</tr>
<tr>
<td>( 0.5 )</td>
<td>-5.56 (13.18)</td>
<td>-5.56 (13.18)</td>
</tr>
<tr>
<td>( 1 )</td>
<td>-1.39 (7.41)</td>
<td>-1.39 (7.41)</td>
</tr>
<tr>
<td>( 1.5 )</td>
<td>-0.09 (2.73)</td>
<td>-0.09 (2.73)</td>
</tr>
<tr>
<td>( 2 )</td>
<td>-0.11 (1.72)</td>
<td>-0.11 (1.72)</td>
</tr>
<tr>
<td>( 3 )</td>
<td>-0.02 (0.56)</td>
<td>-0.02 (0.56)</td>
</tr>
<tr>
<td>( \tau / T = 0.5 )</td>
<td>| |</td>
<td></td>
</tr>
<tr>
<td>( 0.5 )</td>
<td>0.3 (12.32)</td>
<td>0.3 (12.32)</td>
</tr>
<tr>
<td>( 1 )</td>
<td>-0.05 (6.13)</td>
<td>-0.05 (6.13)</td>
</tr>
<tr>
<td>( 1.5 )</td>
<td>0.09 (3.27)</td>
<td>0.09 (3.27)</td>
</tr>
<tr>
<td>( 2 )</td>
<td>0.02 (1.37)</td>
<td>0.02 (1.37)</td>
</tr>
<tr>
<td>( 3 )</td>
<td>0.01 (0.56)</td>
<td>0.01 (0.56)</td>
</tr>
</tbody>
</table>

Table 4: Performance of \( \hat{\tau}_{MLE} \) and \( \hat{\tau}_{LAD} \): \( \tau / T = 0.3 \) and \( 0.5; \tau = 0, 4, 6 \) and \( 9; \delta = 0.5, 1, 1.5, 2 \) and \( 3; \) and \( T = 50 \)

### 6. Conclusions and future work.

In this research a change-point estimator based on robust regression was proposed and its performance was analyzed. Based on the numerical results presented above, we recommend following the next scheme when a series with a change in the mean is suspected to have an atypical observation: if the observation which is suspected to be an outlier is located at the beginning of the series (that is, before the second quantile of the series), then use \( \hat{\tau}_{MLE} \) to estimate the change-point. But, if the outlier is located after the second quantile of the series, we recommend using \( \hat{\tau}_{LAD} \) estimator regardless of the size of the series or the size of the shift in the mean. Also, if the size of the change is greater or equal than 1.5\( \sigma \) LAD estimator has better performance than MLE.

For future work, further simulations for different scenarios will be done in order to compare the numerical bias and standard error of both estimators. Development of an analogue estimator for changes only in variance and in both parameters at same time for series with outliers based on the least square deviations is left as well as a comparison with their MLE’s counterparts; and determine the performance of \( \hat{\tau}_{LAD} \) for series which have multiple shifts in the mean.

### Acknowledgements

We give special thanks to PROMEP (grant F-PROMEP-38/Rev-03 SEP-23-005) and PAICYT (PAICYT IT990-11) for enabling this research. Its commitment to science made this possible. We also appreciate the comments provided by the anonymous ISERC referee. The corresponding suggestions improved the quality of this manuscript.
7. References


